

Proposal for Predicting Movie Box Office Revenue Using Machine Learning

Anant Sahoo, Deep Patel, Sulaiman Mohyuddin, William Douglass, Will Sessoms

Abstract—This project aims to develop a regression model to predict the box office revenue of movies prior to their release. Using the TMDb 5000 Dataset, which contains diverse information about films, including budget, genres, keywords, production companies, and cast, we intend to explore which factors most strongly influence revenue. By analyzing these relationships, we seek to build a predictive model that can provide actionable insights for studios and producers before production or release. The project also considers the development of a simple web interface to demonstrate the model's capabilities in practice. Beyond prediction, the work highlights the value of applying machine learning methods to a real-world problem where uncertainty, missing information, and high financial stakes intersect.

I. INTRODUCTION

A. Objective

The objective of this project is to build a regression model that predicts the box office revenue of movies. The model will use features available before a film's release, such as budget, genres, production company, language, keywords, and key cast members. By focusing on pre-release data, the model will be designed to estimate financial success ahead of time, allowing stakeholders to make informed decisions before production and release.

B. Motivation

Predicting box office revenue has practical value for multiple stakeholders in the film industry. Film production is a high-risk investment: a typical studio film can require tens or even hundreds of millions of dollars in funding before release, and marketing expenses often match or exceed the budget itself. Despite such high costs, the success of a movie is often unpredictable, making financial forecasting an important area for improvement. Being able to estimate the outcome of a project before release can reduce uncertainty for studios and investors. Accurate predictions can help determine whether a project is worth funding, guide marketing and promotion strategies, and even influence casting or directing decisions.

The model may also reveal broader trends in audience preferences across genres, budgets, and production strategies, creating opportunities for studios to align future projects with market demand. For example, studios may discover that certain combinations of genres, release seasons, or production companies consistently yield higher returns. Beyond its practical applications, this project provides a chance to apply machine learning techniques to a real-world dataset, offering both technical experience and industry insight. It also demonstrates how data-driven approaches can complement

traditional decision-making in fields that are highly creative but financially constrained.

II. DATA OVERVIEW

For this project, we use the TMDb 5000 Dataset from Kaggle, which includes information on approximately 5000 films. The dataset is split across two CSV files: one containing credits (title, cast, and crew) and another containing detailed movie statistics, including budget, revenue, popularity, production company, release date, runtime, genres, vote averages, and vote counts.

The credits data requires preprocessing since cast and crew information are stored as JSON strings. These fields will be parsed into structured formats, allowing us to extract key information such as directors and top-billed cast members. Monetary values such as budget and revenue will be adjusted for inflation to ensure comparability across decades of films.

While the dataset is comprehensive, it does contain challenges such as missing values and zero entries in budget and revenue columns. These issues will be addressed through careful data cleaning and filtering. Beyond cleaning, we also plan to create engineered features. For instance, categorical variables such as genre, production company, and release month can be encoded to capture patterns, while cast and crew information can be aggregated into summary statistics about their past successes. Another important step will be identifying and removing or adjusting for outliers, such as films with unusually high or low reported values that may distort model training. By refining the dataset into a structured and reliable format, we can ensure that the features used for modeling are both accurate and representative of real-world conditions.

III. WORK DISTRIBUTION

The responsibilities for this project are distributed among team members as follows:

- Deep Patel: Frontend development and machine learning model development.
- Anant Sahoo: Data collection, data cleaning, and exploratory data analysis.
- William Douglass: Data processing and exploratory data analysis.
- Will Sessoms: Data processing and machine learning model development.
- Sulaiman Mohyuddin: Frontend development, data collection, and data cleaning.

IV. TIMELINE OF MILESTONES

- **Data Collection and Cleaning:** Partly completed using the TMDb 5000 dataset. Future steps may involve integrating additional features such as social media buzz, trailer views, and advertising budgets. This stage ensures that we have a consistent and usable dataset to build on.

Deliverable: A raw dataset in a unified format.

- **Data Processing:** Parse credits JSON into structured formats, extract top actors and directors, adjust financial figures for inflation, and filter out irrelevant or noisy features. Processing will also include the creation of engineered variables to capture hidden relationships in the data.

Deliverable: A clean, processed CSV suitable for modeling.

- **Exploratory Data Analysis (EDA):** Conduct a correlation matrix to identify influential features, explore genre-based trends, detect outliers, and evaluate preliminary model hyperparameters. This stage will also involve creating visualizations that highlight patterns across budget, genre, and cast. The EDA will guide both feature selection and modeling decisions.

Deliverable: An EDA report with visualizations and findings.

- **Machine Learning Model Development:** Begin with baseline regression models such as linear regression, then advance to more complex methods, including Random Forests, AdaBoost, or Neural Networks. Models will be evaluated using cross-validation, and performance will be measured with metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Hyperparameter tuning will be performed to improve model accuracy. A final evaluation will compare the model against simple baselines, such as budget-only predictors, to ensure meaningful improvement.

Deliverable: A refined predictive model with error analysis.

- **Frontend Development:** Create a simple JavaScript interface that accepts movie attributes and returns predicted box office revenue from the trained model. The interface will serve as a proof-of-concept demonstration that makes the model's output more accessible to non-technical users.

Deliverable: A functional web page connected to the ML model.

insights gained from the model can guide decision-making in the film industry.

If time permits, the final deliverable will also include a simple web application demonstrating the model's predictions. Regardless of interface development, the project will provide value by offering a structured, data-driven understanding of what drives movie revenue. Beyond industry use, the project serves as a learning opportunity for applying machine learning techniques to a large, messy dataset, preparing the team for more advanced work in data science and predictive modeling in the future. Looking ahead, the framework could be expanded to integrate additional signals, such as marketing spend or online attention, or even adapted for predicting streaming success as the entertainment industry continues to evolve. In this way, the project not only fulfills its immediate goals but also lays the groundwork for further research and applications.

V. EXPECTED OUTCOME AND CONCLUSIONS

The expected outcome of this project is a regression model capable of predicting box office revenue within a reasonable error margin. Through feature engineering and model development, we aim to highlight which factors—such as budget, genre, and cast—are most predictive of financial success. Even in the absence of a complete user interface, the